Sub-sampled dictionaries for coarse-to-fine sparse representation-based human action recognition

JongHo Lee, Hyun-seok Min, Jeong-jik Seo, Wesley De Neve, and Yong Man Ro

In: IEEE International Conference on Multimedia and Expo (ICME), 2014.

http://ieeexplore.ieee.org/abstract/document/6890317/

SUB-SAMPLED DICTIONARIES FOR COARSE-TO-FINE SPARSE REPRESENTATION-BASED HUMAN ACTION RECOGNITION

JongHo Lee[1,2], *Hyun-seok Min*[1], *Jeong-jik Seo*[1], *Wesley De Neve*[1,3], *Yong Man Ro*[1,*]

[1]Image and Video Systems Lab, Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea
[2]Agency for Defense Development (ADD), Republic of Korea
[3]Multimedia Lab, Ghent University-iMinds, Belgium
{gomdori1976, hsmin, jj.seo, wmdeneve}@kaist.ac.kr, ymro@ee.kaist.ac.kr

ABSTRACT

Automatic human action recognition is a core functionality of systems for video surveillance and human-object interaction. However, the diverse nature of human actions and the noisy nature of most video content make it difficult to achieve effective human action recognition. To overcome the aforementioned problems, Sparse Representation (SR) has recently attracted substantial research attention. However, although SR-based approaches have proven to be reasonably effective, the computational complexity of the testing stage prohibits their usage by applications requiring support for real-time operation and a vast number of human action classes. In this paper, we propose a novel method for human action recognition, leveraging coarse-to-fine sparse representations that have been obtained through dictionary sub-sampling. Comparative experimental results obtained for the UCF50 dataset demonstrate that the proposed method is able to achieve efficient human action recognition, at no substantial loss in recognition accuracy.

Index Terms **—** Coarse-to-fine sparse representation, dictionary sub-sampling, human action recognition.

1. INTRODUCTION

The increasing popularity of video surveillance and human-computer interaction has recently drawn significant research attention to the field of computerized human action recognition. This field aims at automatically understanding human activity in video content [1]. Early techniques for human action recognition focused on capturing, tracking, and analyzing motion [2]. More recently, substantial progress has been made by introducing more descriptive features [3][4] and machine learning-based approaches [5][6]. Despite this progress, vision-based human action recognition remains challenging, mainly due to 1) the highly varying nature of human actions and their context and 2) the noisy nature of most video content.

* Corresponding author: Prof. Yong Man Ro. Tel: +82 42 350 3494.

Sparse Representation (SR) has recently emerged as an effective tool for analyzing audiovisual signals [7]. To that end, SR characterizes the information in an audiovisual signal by means of a linear combination of a small number of discriminative base signals (so-called atoms) taken from a large set of base signals (a so-called dictionary) [8]. SR-based classification was shown to be robust against intra-class variety and noise, and where this robustness can be mainly attributed to the fact that SR uses overcomplete dictionaries [11][12]. The latter means that the number of atoms exceeds the audiovisual signal dimension, so that any audiovisual signal can be represented by more than one combination of different atoms. However, although SR-based approaches come with reasonable levels of recognition accuracy [9][10], the computational complexity of their testing stage may be prohibitive for applications requiring support for real-time operation or a vast number of human action classes [13].

To mitigate the computational complexity of SR-based human action recognition, we propose a novel two-step method. In the first step, we select candidate human action classes for a given test video clip through the use of a dictionary with a limited number of atoms per human action class supported, and where the dimension of the atoms used has also been restricted. In the second step, we classify the human action shown in the test video clip through the use of a dictionary that only contains the candidate human action classes selected, and where these classes are described by atoms that are higher in number and dimension than the atoms used to describe the classes during the first step.

To investigate the feasibility of the proposed method for human action recognition, we performed experiments with the UCF50 dataset [14]. This dataset, which is one of the largest action recognition datasets publicly available, represents a natural pool of various human actions, featuring a wide range of scenes and viewpoints. Our experimental results demonstrate that the proposed method facilitates efficient SR-based human action recognition, at no substantial loss in recognition accuracy.

We organized the remainder of this paper as follows. In Section 2, we briefly review human action recognition, paying particular attention to conventional SR-based

classification. In Section 3, we introduce our SR-based solution towards the problem of human action recognition. In Section 4 and Section 5, we discuss our experimental setup and results, respectively. Finally, we present our conclusions and directions for future research in Section 6.

2. RELATED WORK

In this section, we first review a number of representative research efforts in the area of human action recognition. Next, we explain conventional SR-based human action recognition. Note that we would like to refer the interested reader to [1] and [2] for more in-depth surveys of human action recognition.

2.1. Human action recognition

Human action recognition is the process of labeling human behavior with one or more pre-defined action classes. To that end, most approaches follow a two-step procedure: 1) feature extraction and representation and 2) action detection and classification [2].

The first step aims at characterizing human action volumes with discriminative features. Similar to the authors of [2], we can make a distinction between global features and local features. Global features include space-time features and frequency-domain features. Space-time features allow capturing space and time relationships. These features are typically derived by concatenating the consecutive silhouettes of objects along the time axis [15]. Frequency-domain features can for instance be used to capture the spatial variation of intensity values. Compared to global features, local features such as Scale-Invariant Feature Transform (SIFT; [16]) and Histograms of Oriented Gradients (HOG; [17]) tend to be more robust against noise, occlusions, viewpoint variation, and changes in scale, and where this higher robustness typically comes at a higher computational cost. Besides global and local features, methods have also been proposed that directly or indirectly model the human body in 2-D or 3-D [18].

The second step aims at action detection (localization) and classification, typically by making use of the features extracted during the first step. To that end, generative models such as Hierarchical Markov Models (HMM) and Dynamic Bayesian Networks (DBN) can be used, as well as discriminative models like Support Vector Machines (SVM; [19]), Artificial Neural Networks (ANN; [6]), and SR-based Classification (SRC; [10]).

As previously pointed out in Section 1, despite the success of recent feature extraction and classification technology, vision-based human action recognition is still a challenging issue.

2.2. SR-based human action recognition

Conventional SR characterizes the content of a test video clip $\mathbf{V}$ by means of a linear combination of atoms taken from an overcomplete dictionary $\mathbf{D_o}$, and where this dictionary has been constructed by concatenating the sets of atoms used to represent the different human action classes:

$$\mathbf{D}_o = \left[\mathbf{\Phi}_{o,1} \mid ... \mid \mathbf{\Phi}_{o,k} \mid ... \mid \mathbf{\Phi}_{o,K}\right] \in \Re^{d_o \times N_o}, \quad (1)$$

where $\mathbf{\Phi}_{o,k}$ denotes the set of atoms used to represent the k^{th} human action class, and where d_o and N_o denote the dimension of an atom and the total number of atoms in the dictionary, respectively. Note that N_o is the number of atoms per human action class l_o times the number of human action classes K (i.e., $N_o = l_o \times K$). Further, we can define the set of atoms $\mathbf{\Phi}_{o,k}$ as follows:

$$\mathbf{\Phi}_k = \left[\mathbf{z}_1^k, ..., \mathbf{z}_{l_o}^k\right] \in \Re^{d_o \times l_o}, \quad (2)$$

where $\mathbf{z}_j^i$ denotes the feature vector of the j^{th} training video clip of the i^{th} human action class. Note that general-propose procedures for atom extraction and dictionary construction can be found in [20].

Given an overcomplete dictionary $\mathbf{D_o}$, we can represent the feature vector $\mathbf{y}$ of the test video clip $\mathbf{V}$ as follows [10]:

$$\mathbf{y} \approx \mathbf{D}_o \mathbf{x}_o \in \Re^{d_o}, \quad (3)$$

where $\mathbf{x}_o$ denotes a sparse coefficient vector.

Given the sparse solution $\mathbf{x}_o$ of (3), we can calculate the residual error for each human action as follows:

$$r_k(\mathbf{y}) = \left\| \mathbf{y} - \mathbf{D}_o \delta_k(\mathbf{x}_o) \right\|_1, \quad (4)$$

where r_k is the residual error of the k^{th} human action class and where $\delta_k(\mathbf{x}_o)$ is a new vector whose only nonzero entries are the entries in $\mathbf{x}_o$ associated with the k^{th} human action class. We can subsequently label $\mathbf{V}$ with the human action class that comes with the smallest residual error.

The sparse solution $\mathbf{x}_o$ of (3) is typically determined through the use of l_1-norm minimization, an optimization problem that has been well-studied in the past, resulting in algorithms such as Orthogonal Matching Pursuit [21], Gradient Projection [22], and Homotopy [23]. However, the time complexity of the aforementioned techniques may be prohibitive for applications requiring support for real-time operation and a vast number of human action classes [13]. As an example, the time complexity of homotopy-based techniques can be approximated as follows [13]:

$$O(d_o^2 + d_o N_o). \quad (5)$$

As shown by (5), the time complexity closely depends on the atom dimension and the dictionary size. Therefore, we can mitigate the time complexity by reducing the atom dimension and the dictionary size. However, when doing so, we need to take into account that a trade-off exists between the time complexity and the accuracy of human action recognition.

3. PROPOSED APPROACH

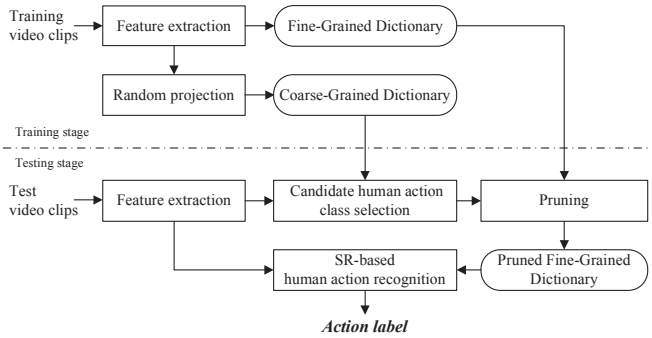

Fig. 1. Proposed approach for SR-based human action recognition.

Fig. 1 shows that the proposed approach for SR-based human action recognition consists of two stages: 1) a training stage and 2) a testing stage.

During the training stage, we generate two dictionaries that are used during testing. Similar to conventional SR (see Section 2.2), we first construct an overcomplete Fine-Grained Dictionary (FGD), containing a set of feature vectors (atoms) extracted from training video clips for each human action class supported. Next, we make use of the FGD to construct a Coarse-Grained Dictionary (CGD), containing a limited number of atoms for each human action class supported, and where the dimension of the atoms used has also been restricted.

During the testing stage, we first make use of the CGD to select candidate human action classes for a given test video clip. We subsequently leverage the candidate human action classes selected to prune the FGD, removing the non-selected human action classes. Finally, we make use of the pruned FGD to classify the human action shown in the test video clip under consideration.

3.1. Dictionary construction

As pointed out in Section 2, a conventional dictionary for SR-based human action recognition may contain a substantial number of atoms for a significant number of human action classes, making SR-based classification inefficient. In addition, we can point out that only a few human action classes are typically related to the human action shown in a given test video clip. Therefore, to reduce the computational complexity of SR-based classification, we first select a few human action classes that are related to the

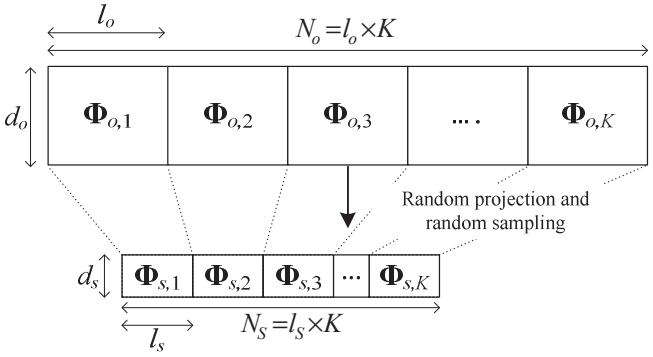

Fig. 2. CGD constructed from FGD.

human action shown in the test video clip under consideration by making use of a sub-sampled dictionary. We refer to this step as candidate human action class selection and then classify the human action shown in the test video clip under consideration through the use of a dictionary that only contains fine-grained information about the selected candidate human action classes. To that end, the proposed method makes use of two dictionaries that are different in size: a Fine-Grained Dictionary (FGD) and a Coarse-Grained Dictionary (CGD). The CGD is a sub-sampled version of the FGD, targeting the selection of candidate human action classes, whereas the FGD is similar to a dictionary used by conventional SR-based classification. Thanks to the fact that the CGD is smaller in size than the FGD, we can efficiently select candidate human action classes for a given test video clip. Next, we can efficiently classify the human action shown in the test video clip by making use of a pruned version of the FGD, only retaining the human action classes selected through the use of the CGD.As shown by Fig. 2, the CGD is a sub-sampled version of the FGD, reducing the size of the FGD by means of random projection [21] (for reducing the atom dimension) and random sampling (for reducing the number of atoms). We can define the CGD as follows:

$$\mathbf{D}_S = \left[\boldsymbol{\Phi}_{S,1} \,|\, ... \,|\, \boldsymbol{\Phi}_{S,k} \,|\, ... \,|\, \boldsymbol{\Phi}_{S,K}\right] \in \Re^{d_S \times l_S \times K}, \tag{6}$$

where $\boldsymbol{\Phi}_{S,k}$ denotes the set of atoms used to characterize the k^{th} human action class.

3.2. Selection of candidate human action classes and human action recognition

During testing, we first select candidate human action classes. Given the feature vector $\mathbf{y}$ of a test video clip $\mathbf{V}$, the reduced feature vector $\mathbf{y}_s$ can be represented as follows:

$$\mathbf{y}_s \approx \mathbf{D}_s \mathbf{x}_s \in \Re^{d_s}, \tag{7}$$

where $\mathbf{x}_s$ denotes a sparse coefficient vector.

Given the sparse solution $\mathbf{x}_s$ of (7), we can calculate the

concentration of the sparse coefficients for each human action class as follows [24]:

$$\mathrm{SCC}_{\mathrm{k}}(\mathrm{x}_{\mathrm{s}}) = \frac{\left\|\delta_{\mathrm{k}}(\mathrm{x}_{\mathrm{s}})\right\|_1}{\left\|\mathrm{x}_{\mathrm{s}}\right\|_1}, \quad (8)$$

where SCC_k denotes the concentration of the sparse coefficients obtained for the k^{th} human action class. Consequently, we can sort all human action classes in descending order by SCC_k. We can then select the candidate human action classes by taking the upper H human action classes. This means that the other human action classes in the FGD can be ignored. As a result, we prune the FGD by removing the non-selected candidate human action classes, resulting in the dictionary $\mathbf{D}_{pr}$:

$$\mathbf{D}_{pr} = \left[\boldsymbol{\Phi}_{pr,1} \mid \ldots \mid \boldsymbol{\Phi}_{pr,h} \mid \ldots \mid \boldsymbol{\Phi}_{pr,H}\right] \in \Re^{d_o \times l_o \times H}. \quad (9)$$

Given the overcomplete dictionary $\mathbf{D}_{pr}$, we can then represent the feature vector $\mathbf{y}$ of a test video clip $\mathbf{V}$ as follows:

$$\mathbf{y} \approx \mathbf{D}_{pr}\mathbf{x}_{pr} \in \Re^{d_o}, \quad (10)$$

where $\mathbf{x}_{pr}$ contains the sparse coefficients obtained for the feature vector $\mathbf{y}$ of the test video clip $\mathbf{V}$. We can subsequently label $\mathbf{V}$ with the candidate human action class that comes with the smallest residual error.

3.3. Time complexity of the proposed method

As described in Section 3.2, the proposed method computes two distinct sparse representations during testing: 1) a coarse sparse representation for candidate human action class selection and 2) a fine sparse representation for identifying the human action class shown in the given test video clip among the selected candidate human action classes. The dictionary used during the first step has a size of $d_s \times (l_s \times K)$, whereas the dictionary used during the second step has a size of $d_o \times (l_o \times H)$. As such, we can approximate the time complexity of the proposed method as follows:

$$O(d_S^2 + d_S l_S K) + O(d_o^2 + d_o l_o H). \quad (11)$$

Given that H is always smaller than K, the time complexity described by (11) is lower than the time complexity described by (5). As a result, we can conclude that the time complexity of the proposed method is lower than the time complexity of conventional SR-based classification (using the FGD).

4. EXPERIMENTAL SETUP

The main purpose of our experiments is to evaluate the proposed method in terms of effectiveness and time complexity. To study the effectiveness of the proposed method, we conducted experiments with the UCF50 dataset, which is one of the largest action recognition datasets publicly available [14]. This dataset covers 50 different types of actions (i.e., $K = 50$), containing 6681 real-world video clips in total. For all the 50 action classes (e.g., biking, golf swing, playing piano, and so on), the video clips in each action class are grouped into 25 groups.

In order to make sure that the reported results are consistent, we obtained our experimental results by making use of "Leave-One-Group-Out Cross-Validation" [14] to train and test both the conventional SR-based method and the proposed method. In particular, we used one group for testing purposes and the remaining 24 groups for training purposes. We then repeated the aforementioned approach in such a way that each group of each action class is used once for testing purposes. Finally, we computed the effectiveness of testing by averaging the 25 results obtained.

To extract atoms from the given training video clips, we used the Cuboid detector to find local keypoints and HOG to describe the local keypoints found, a combination that is known to lead to high recognition rates [20]. The Cuboid detector relies on separable linear filters for computing the response function of a video clip $\mathbf{V}(x,y,t)$. The response function computed is of the form $R=(\mathbf{V}\times g\times h_{ev})^2+(\mathbf{V}\times g\times h_{od})^2$, where $g(x,y;\sigma)$ is a 2-D Gaussian smoothing function, applied only along the spatial dimensions, and where h_{ev} and h_{od} are quadrature pairs of 1-D Gabor filters. These quadrature pairs are defined as $h_{ev}(t;\tau,\omega) = -\cos(2\pi t\omega)\exp^{-t^2/\tau^2}$ and $h_{od}(t;\tau,\omega) = -\sin(2\pi t\omega)\exp^{-t^2/\tau^2}$, with $\omega=4/\tau$. The parameters σ and τ roughly correspond to the spatial and temporal scales used by the Cuboid detector [20]. In our experiments, we used values of 4 and 2 for σ and τ, respectively.

To describe local keypoints, we adopted HOG with a dimension of 1,440 [20]. Due to the high dimensionality of these histograms, we projected the descriptors generated on a random 144-D space (i.e., $d_o = 144$), given that projection on a random lower dimensional subspace is able to reliably preserve vector distance [21]. Further, to generate the FGD, we randomly selected 3000 descriptors (i.e., $l_o = 3{,}000$) for each human action class (i.e., the total number of atoms in the dictionary is 150,000). In addition, for pruning the FGD, we set the number of candidate human action classes to 10 (i.e., $H = 10$).

We used homotopy-based l_1-norm minimization, setting the number of iterations to 5000 and the error tolerance to 0.5. To measure the effectiveness of proposed approach, we made use of accuracy:

$$Accuracy = \frac{N_{true}}{N_{total}}, \quad (12)$$

where N_{total} and N_{true} denote the total number of test video clips and the number of test video clips labeled with the true action, respectively.

5. EXPERIMENTAL RESULTS

To investigate the feasibility of the proposed method for human action recognition, we compared its effectiveness and efficiency with the effectiveness and efficiency of conventional SR-based human action recognition. To that end, we first varied the atom dimension d_o from 48 to 144 by making use of random projection [25]. Furthermore, we also varied the numbers of atoms per human action class l_o by making use of random sampling (150, 300, 600, 900, and 1500). In contrast, the conventional method uses fixed values for the aforementioned parameters (i.e., $d_o = 144$ and $l_o = 1500$, respectively). Next, we selected candidate human action classes via coarse SR-based classification using the different CGDs constructed. Finally, we performed fine SR-based classification of the human actions shown in the test video clips used.

Fig. 3 shows the recognition accuracy obtained for the different human action recognition approaches. The x-axis represents the number of atoms used per human action class, whereas the y-axis denotes the recognition accuracy.

We can observe that the proposed method allows for efficient human action recognition at no significant loss in recognition accuracy (i.e., the loss is less than 11%). Specifically, the recognition accuracy of the proposed method is higher than the recognition accuracy of the conventional method when using more than 600 atoms per human action class. This means that the true human action class is then typically among the selected candidate human action classes, even when making use of a pruned FGD.

We can also observe that the recognition accuracy of the proposed method is relatively robust against changes in the atom dimension and the number of atoms used per human action class. This is thanks to the use of a pruned FGD that is still overcomplete. Given the pruned FGD, we can classify the human action shown in the test video clip, even when the true human action class does not come with the smallest residual error during candidate human action class selection.

Fig. 4 shows the time complexity obtained for the different human action recognition approaches. Compared to the time complexity of the conventional method, we can observe that the time complexity of the proposed method is approximately two times lower, for the different parameter settings used.

6. CONCLUSIONS AND FUTURE RESEARCH

In this paper, we proposed a novel method for human action recognition, leveraging coarse-to-fine sparse representations that have been obtained through dictionary sub-sampling. Specifically, to select candidate human action classes, we first perform SR-based human action recognition using a coarse-grained dictionary. We then classify the human action shown in the test video clips using a fine-grained dictionary that only contains the candidate human action classes selected. As shown by comparative experimental results obtained for the UCF50 dataset, the proposed method is able to achieve efficient human action recognition, at no substantial loss in recognition accuracy.

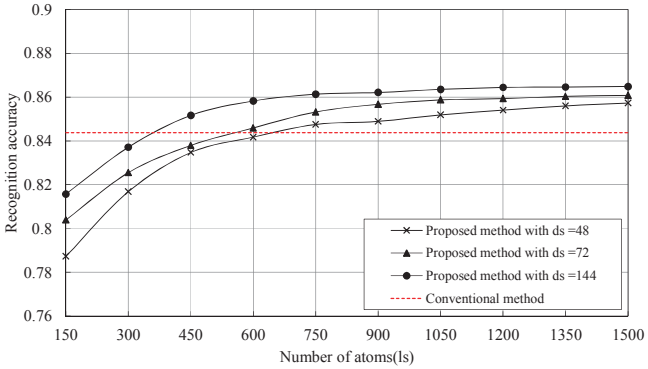

Fig. 3. Accuracy of different human action recognition approaches.

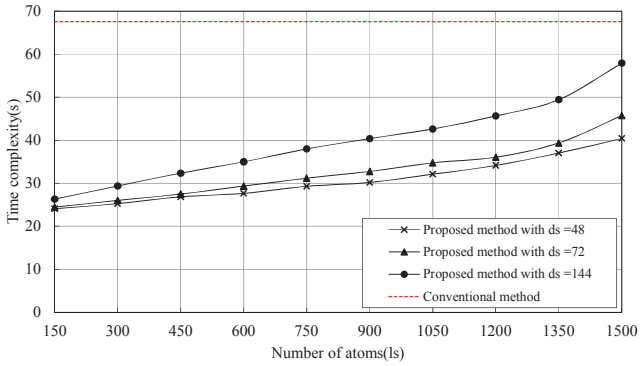

Fig. 4. Time complexity of different human action recognition approaches.

We can identify a number of directions for future research. First, we plan to investigate more sophisticated approaches for generating the sub-sampled dictionaries. Second, we plan to compare the efficiency and effectiveness of the proposed method with other state-of-the-art techniques, hereby also making use of other publicly available datasets.

7. ACKNOWLEDGEMENTS

This work was supported under the framework of international cooperation program managed by National Research Foundation of Korea (NRF-2012K2A1A2033054)

8. REFERENCES

[1] J. K. Aggarwal and M. S. Ryoo, "Human Activity Analysis: A Review," *ACM Computing Surveys*, vol. 43, no. 3, 2011.

[2] S. R. Ke, H. L. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, "A Review on Video-based Human Activity Recognition" *Computers 2013*, vol. 2, no. 2, pp. 88-131, 2013.

[3] A. A. Oikonomopoulos, I. Patras, and M. Pantic. "Spatio-temporal salient points for visual recognition of human actions," *IEEE Trans. Systems, Man, and Cybernetics, Part B*, vol. 36, no. 3, pp. 710–719, 2006.

[4] G. Willems, T. Tuytelaars, and L. Van Gool. "An efficient dense and scale-invariant spatio-temporal interest point detector," *European Conf. on Computer Vision*, pp. 1-14, 2008.

[5] V. Vapnik, S. E. Golowich, and A. Smola, "Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing," *Adv. Neural Inf. Process. Syst.*, vol. 9, pp. 281–287, 1997.

[6] M. K. Fiaz and B. Ijaz, "Vision based Human Activity Tracking using Artificial Neural Networks," *IEEE Int'l Conf. on Intelligent and Advanced Systems*, pp. 1–5, 2010.

[7] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, "Sparse Representation For Computer Vision and Pattern Recognition," *Proceeding of IEEE*, vol. 88, no. 6, pp.1031-1044, 2009.

[8] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation," *IEEE Trans. on Signal Processing*, vol. 54, No 11, pp. 4311-4322, 2006.

[9] K. Huang and S. Aviyente, "Sparse Representation for Signal Classification," *In Adv. NIPS*, 2006.

[10] T. Guha and W. R. Kreidieh, "Learning Sparse representations for Human Action Recognition," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 34, no. 8, pp. 1576-1588, 2012.

[11] B. Wohlberg, "Noise Sensitivity of Sparse Signal Representations: Reconstruction Error Bounds for the Inverse Problem," *IEEE Trans. Signal Processing*, vol. 51, no. 12, pp. 3053-3060, 2003.

[12] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210-227, 2009.

[13] A. Y. Yang, A. Ganesh, Z. H. Zhou, S. S. Sastry, and Y. Ma, "Fast l1-minimization algorithms and application in robust face recognition," *UC Berkeley, Tech. Rep.*, pp. 1-12, 2010.

[14] K. R. Kishore and M. Shah, "Recognizing 50 Human Action Categories of Web Videos," *Machine Vision and Applications Journal (MVAP)*, vol. 24, no. 5, pp. 971-981, 2012.

[15] Y. Ke, R. Sukthankar, and M. Hebert, "Spatio-temporal Shape and Flow Correlation for Action Recognition," *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1-8, 2007.

[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91-100, 2004.

[17] N. Dalal and B.Triggs, "Histograms of Oriented Gradients for Human Detection," *IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 1, pp. 886-893, 2005.

[18] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 780–785, 1997.

[19] V. Vapnik, S.E. Golowich, and A. Smola, "Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing," *Adv. Neural Inf. Process. Syst.*, vol. 9, pp. 281–287. 1997.

[20] H. Wang, M.M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of Local Spatio-Temporal Features for Action Recognition," *British Machine Vision Conf.*, 2009.

[21] G. Davis, S. Mallat, "Adaptive greedy approximations". *Journal of Constructive Approximation*, 13:57–98, 1997.

[22] R. Gribonval, M. Nielsen, "Sparse representations in unions of bases". *IEEE Trans. Info. Theory*, 49(12):1320–1325, 2003.

[23] D. Malioutov, M. Cetin, and A. Willsky, "Homotopy continuation for sparse signal representation". *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005.

[24] S. H. Lee, H. Kim, K. N. Plataniotis, and Y. M. Ro, "Using Color Texture Sparsity for Facial Expression Recognition," *IEEE Int'l Conf. on Automatic Face and Gesture Recognition* (FG), pp. 1-6, 2013.

[25] R. Baraniuk and M. Wakin, "Random Projections of Smooth Manifolds," *Foundations of Computational Math.*, vol. 9, pp. 51-77, 2009.